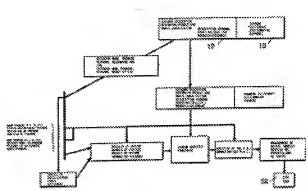
Method and apparatus for speech synthesis by wave form overlapping and adding.

Publication number: EP0363233
Publication date: 1990-04-11
Inventor: HAMON CHRISTIAN
Applicant: FRANCE ETAT (FR)
Classification:
- international: ***G10L19/00; G10L13/00; G10L13/06;*** **G10L19/00; G10L13/00;** (IPC1-7): G10L5/04
- European: G10L13/06C
Application number: EP19890402394 19890901
Priority number(s): FR19880011517 19880902

Also published as:
WO9003027 (A1)
US5327498 (A1)
FR2636163 (A1)
DK107390 (A)
EP0363233 (B1)

Abstract of **EP0363233**

Speech is synthesised from elements such as diphones. At least the vocalized sounds of the sound elements are subjected to window analysis centred essentially on the start of each pulsed response of the vocal tract to the stimulation of the vocal cords, using a filtering window whose amplitude decreases to zero at the edges of the window and whose width is equal to at least twice the fundamental period of origin or twice the fundamental period of synthesis. The signals corresponding to each sound element which are obtained from the window analysis are replaced, the phase shift of the signals being equal to the fundamental period of synthesis, which is greater or less than the fundamental period of origin, depending on the prosodic data concerning the synthesis frequency. Synthesis is carried out by addition of the phase-shifted signals.



(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Numéro de publication: **0 363 233**
A1

(12) DEMANDE DE BREVET EUROPEEN

(21) Numéro de dépôt: 89402394.4

(51) Int. Cl.5: G10L 5/04

(22) Date de dépôt: 01.09.89

(30) Priorité: 02.09.88 FR 8811517

(43) Date de publication de la demande:
11.04.90 Bulletin 90/15

(84) Etats contractants désignés:
BE DE ES GB IT NL SE

(71) Demandeur: ETAT FRANCAIS représenté par le Ministre des Postes, des Télécommunications et de l'Espace (CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS) 38-40 rue du Général Leclerc
F-92131 Issy-les-Moulineaux(FR)

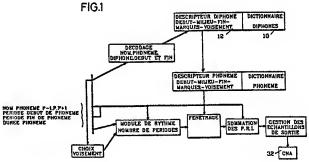
(72) Inventeur: Hamon, Christian
Le Grand Clos Quevert
F-22100 Dinan(FR)

(74) Mandataire: Fort, Jacques
CABINET PLASSERAUD 84, rue d'Amsterdam
F-75009 Paris(FR)

(54) Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'onde.

(57) La synthèse de parole est effectuée à partir d'éléments tels que des diphones. On effectue, au moins sur les sons voisés des éléments sonores, une analyse par fenêtrage sensiblement centrée sur le début de chaque réponse impulsionnelle du conduit vocal à l'excitation des cordes vocales à l'aide d'une fenêtre de filtrage présentant une amplitude décroissant jusqu'à zéro aux bords de la fenêtre de largeur au moins égale à deux fois la période fondamentale d'origine ou deux fois la période fondamentale de synthèse, on replace les signaux résultant du fenêtrage correspondant à chaque élément sonore, avec un décalage temporel de ceux-ci égal à la période fondamentale de synthèse, inférieure ou supérieure à la période fondamentale d'origine, suivant l'information prosodique concernant la fréquence de synthèse, on effectue la synthèse par sommation des signaux ainsi décalés.

FIG.1



EP 0 363 233 A1

Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'onde.

L'invention concerne les procédés et dispositifs de synthèse de la parole ; elle concerne, plus particulièrement, la synthèse à partir d'un dictionnaire d'éléments sonores par découpage du texte à synthétiser en microtrames identifiées chacune par un numéro d'ordre d'élément sonore correspondant et par des paramètres prosodiques (information de hauteur de son au début et à la fin de l'élément sonore et durée de l'élément sonore), puis par adaptation et concaténation des éléments sonores par une procédure d'addition-recouvrement.

Les éléments sonores stockés dans le dictionnaire seront fréquemment des diphones, c'est-à-dire des transitions entre phonèmes, ce qui permet, pour la langue française, de se contenter d'un dictionnaire d'environ 1300 éléments sonores ; on peut cependant utiliser des éléments sonores différents, par exemple des syllabes ou même des mots. Les paramètres prosodiques sont déterminés en fonction de critères portant sur le contexte : la hauteur de son qui correspond à l'intonation, dépend de l'emplacement de l'élément sonore dans un mot et dans la phrase et la durée donnée à l'élément sonore est fonction du rythme de la phrase.

Il faut rappeler au passage que les méthodes de synthèse de la parole se subdivisent en deux groupes. Celles qui utilisent un modèle mathématique du conduit vocal (synthèse par prédiction linéaire, synthèse à formants et synthèse à transformée de Fourier rapide) font intervenir une déconvolution de la source et de la fonction de transfert du conduit vocal et exigent en général une cinquantaine d'opérations arithmétiques par échantillon numérique de la parole avant conversion numérique-analogique et restitution.

Cette déconvolution source-conduit vocal permet d'une part la modification de la valeur de la fréquence fondamentale des sons voisés, c'est-à-dire des sons qui ont une structure harmonique et sont provoqués par vibration des cordes vocales, et d'autre part la compression des données représentant le signal de parole.

Celles qui appartiennent au second groupe de procédés utilisent la synthèse dans le domaine temporel par concaténation de formes d'onde. Cette solution a l'avantage de la flexibilité d'emploi et de la possibilité de réduire considérablement le nombre d'opérations arithmétiques par échantillons. En contrepartie, elle ne permet pas de réduire le débit nécessaire à la transmission autant que les méthodes basées sur un modèle mathématique. Mais cet inconvénient disparaît lorsqu'on recherche essentiellement une bonne qualité de restitution sans être gêné par la nécessité de transmettre des données sur un canal étroit.

La synthèse de parole suivant la présente invention appartient au second groupe. Elle trouve une application particulièrement importante dans le domaine de la transformation d'une chaîne orthographique (constituée par exemple par le texte fourni par une imprimante) en un signal de parole, par exemple restitué directement ou émis sur une ligne téléphonique normale.

On connait déjà (Diphone synthesis using an overlap-add technique for speech waveforms concatenation, CHARPENTIER et al, ICASSP 1986, IEEE-IECEJ-ASJ International Conference on Acoustics Speech and Signal Processing, pages 2 015-2 018)un procédé de synthèse de parole à partir d'éléments sonores utilisant une technique d'addition-recouvrement de signaux à court-terme. Mais il s'agit de signaux à court-terme de synthèse, avec normalisation du recouvrement des fenêtres de synthèse, obtenus par un processus très complexe :

- analyse du signal original par fenêtrage synchrone du voisement ;
- transformée de Fourier du signal à court-terme ;
- détection d'enveloppe ;
- homothétie de l'axe fréquentiel sur le spectre de la source ;
- pondération du spectre modifié de la source par l'enveloppe du signal d'origine ;
- transformée de Fourier inverse.

La présente invention vise notamment à fournir un procédé relativement simple et permettant une reproduction acceptable de la parole. Elle part de l'hypothèse qu'on peut considérer les sons voisés comme la somme des réponses impulsionnelles d'un filtre, stationnaire durant plusieurs millisecondes, (correspondant au conduit vocal) excité par une suite de Dirac, c'est-à-dire par un "peigne d'impulsions", de façon synchrone de la fréquence fondamentale de la source, c'est-à-dire des cordes vocales, ce qui se traduit dans le domaine spectral par un spectre harmonique, les harmoniques étant espacés de la fréquence fondamentale et pondérés par une enveloppe présentant des maxima appelés formants, dépendant de la fonction de transfert du conduit vocal.

On a déjà proposé (Micro-phonemic method of speech synthesis, Lucaszewic et al, ICASSP 1987, IEEE, pages 1426-1429) d'effectuer une synthèse de parole où la diminution de la fréquence fondamentale des sons voisés, lorsqu'elle est nécessaire pour respecter des données prosodiques, est effectuée par insertion de zéros, les microphonèmes stockés devant alors obligatoirement correspondre à la hauteur maximale possible du son à restituer,

ou bien (brevet US 4 692 941) de diminuer de la même manière par insertion de zéros la fréquence fondamentale, et d'augmenter celle-ci en diminuant la taille de chaque période. Ces deux méthodes introduisent sur le signal de parole des distorsions non négligeables lors de la modification de la fréquence fondamentale.

La présente invention vise à fournir un procédé et un dispositif de synthèse à concaténation de formes d'onde ne présentant pas la limitation ci-dessus et permettant de fournir une parole de bonne qualité, tout en ne nécessitant qu'un faible volume de calculs arithmétiques.

Dans ce but, l'invention propose notamment un procédé caractérisé en ce que :

- on effectue, au moins sur les sons voisés des éléments sonores, un fenêtrage centré sur le début de chaque réponse impulsionnelle du conduit vocal à l'excitation des cordes vocales (ce début pouvant être mémorisé dans un dictionnaire) à l'aide d'une fenêtre présentant un maximum pour ledit début et une amplitude décroissant jusqu'à zéro au bord de la fenêtre, et
- on replace les signaux fenêtrés correspondant à chaque élément sonore avec un décalage temporel égal à la période fondamentale de synthèse à obtenir, inférieur ou supérieur à la période fondamentale d'origine suivant l'information prosodique de hauteur de la fréquence fondamentale et on effectue une sommation de ces signaux.

Ces opérations constituent la procédure de recouvrement puis addition des formes d'onde élémentaires obtenues par fenêtrage du signal de parole.

En général, on utilisera des éléments sonores constitués par des diphones.

La largeur de la fenêtre peut varier entre des valeurs inférieures et supérieures à deux fois la période d'origine. Dans l'exemple de mise en oeuvre qui sera décrit plus loin, la largeur de la fenêtre est choisie avantageusement égale à environ deux fois la période d'origine en cas d'augmentation de la période fondamentale ou environ deux fois la période finale de synthèse en cas d'augmentation de la fréquence fondamentale, afin de compenser partiellement les modifications d'énergie dues au changement de la fréquence fondamentale, non compensées par une normalisation possible de l'énergie, tenant compte de la contribution de chaque fenêtre à l'amplitude des échantillons du signal numérique de synthèse : dans le cas d'une diminution de la période fondamentale, la largeur de la fenêtre sera donc inférieure à deux fois la période fondamentale d'origine. Il est peu souhaitable de descendre au dessous de cette valeur.

Du fait qu'il est possible de modifier la valeur de la fréquence fondamentale dans les deux sens, les diphones sont mémorisés avec la fréquence fondamentale naturelle du locuteur.

Avec une fenêtre de durée égale à deux périodes fondamentales consécutives dans le cas voisé, on obtient des formes d'onde élémentaires dont le spectre représente sensiblement l'enveloppe du spectre du signal de parole ou spectre à court terme large bande -du fait que ce spectre est obtenu par convolution du spectre harmonique du signal de parole et de la réponse fréquentielle de la fenêtre, qui dans ce cas possède une largeur de bande supérieure à la distance entre harmoniques- ; la redistribution temporelle de ces formes d'onde élémentaires donnera un signal possédant sensiblement la même enveloppe que le signal d'origine mais une distance entre harmoniques modifiée.

Avec une fenêtre de durée supérieure à deux périodes fondamentales, on obtient des formes d'onde élémentaires dont le spectre est encore harmonique, ou spectre à court terme bande étroite -du fait que cette fois-ci la réponse fréquentielle de la fenêtre est moins large que la distance entre harmoniques- ; la redistribution temporelle de ces formes d'onde élémentaires donnera un signal possédant, comme le signal de synthèse précédent, sensiblement la même enveloppe que le signal d'origine à ceci près qu'on aura introduit des termes de réverbération (signaux dont le spectre possède une amplitude moindre, une phase différente, mais la même forme que le spectre d'amplitude du signal d'origine), dont l'effet ne sera audible qu'au delà de largeurs de fenêtre d'environ trois périodes, cet effet de réverbération ne dégradant pas la qualité du signal de synthèse lorsque son amplitude est faible.

On peut notamment utiliser une fenêtre de Hanning, bien que d'autres formes de fenêtre soient également acceptables.

Le traitement défini ci-dessus peut également être appliqué aux sons dits sourds ou non voisés, pouvant être représentés par un signal dont la forme s'apparente à celle d'un bruit blanc, mais sans synchronisation des signaux fenêtrés : ceci a pour but d'homogénéiser le traitement sur les sons sourds et les sons voisés, ce qui permet d'une part le lissage entre éléments sonores (diphones) et entre phonèmes sourds et voisés, et d'autre part une modification du rythme. Il se pose un problème à la jonction entre diphones. Une solution pour écarter cette difficulté consiste à omettre l'extraction de formes d'onde élémentaires à partir des deux périodes fondamentales adjacentes de transition entre diphones (dans le cas des sons sourds, les marques de voisement sont remplacées par des marques posées arbitrairement) : on pourra soit définir une troisième fonction d'onde élémentaire en calculant la moyenne des deux fonctions d'onde élémentaires extraites de part et d'autre du diphone, soit utiliser la procédure d'addition-recou-

vrement directement sur ces deux fonctions d'onde élémentaires.

L'invention sera mieux comprise à la lecture de la description qui suit d'un mode particulier de mise en oeuvre de l'invention, donné à titre d'exemple non limitatif. La description se réfère aux dessins qui l'accompagnent, dans lesquels :

- la Figure 1 est un graphe destiné à illustrer la synthèse de la parole par concaténation de diphones et modification des paramètres prosodiques dans le domaine temporel, conformément à l'invention ;

- la Figure 2 est un schéma synoptique montrant une constitution possible du dispositif de synthèse, implanté sur un calculateur hôte ;

- la Figure 3 montre, à titre d'exemple, comment on modifie les paramètres prosodiques d'un signal naturel, dans le cas d'un phonème particulier ;

- les Figures 4A, 4B et 4C sont des graphiques destinés à montrer des modifications spectrales apportées à des signaux de synthèse voisés, la Figure 4A montrant le spectre d'origine, la Figure 4B le spectre avec diminution de la fréquence fondamentale et la Figure 4C le spectre avec augmentation de cette fréquence ;

- la Figure 5 est un graphique montrant un principe d'atténuation des discontinuités entre diphones ;

- la Figure 6 est un schéma montrant le fenêtrage sur plus de deux périodes.

La synthèse d'un phonème est effectuée à partir de deux diphones stockés dans un dictionnaire, chaque phonème étant composé de deux demi-diphones. Le son "é" dans "période" par exemple sera obtenu à partir du second demi-diphone de "pai" et du premier demi-diphone de "air".

Un module de traduction orthographique phonétique et de calcul de la prosodie (qui ne fait pas partie de l'invention) fournit à un instant donné, des indications identifiant :
- le phonème à restituer, d'ordre P
- le phonème précédent, d'ordre P-1
- le phonème suivant, d'ordre P + 1
et donnant la durée à affecter au phonème P ainsi que les périodes au début et à la fin (Figure 1).

Une première opération d'analyse, qui n'est pas modifiée par l'invention, consiste à déterminer, par décodage du nom des phonèmes et des indications prosodiques, les deux diphones retenus pour le phonème à utiliser et le voisement.

Tous les diphones disponibles (au nombre de 1300 par exemple) sont mémorisés dans un dictionnaire 10 muni d'une table constituant le descripteur 12 et contenant l'adresse du début de chaque diphone (en nombre de blocs de 256 octets) la longueur du diphone et le milieu du diphone (ces deux derniers paramètres étant exprimés en nombre d'échantillons à partir du début) et des marques de voisement repérant le début de la réponse du conduit vocal à l'excitation des cordes vocales dans le cas d'un son voisé (au nombre de 35 par exemple). Des dictionnaires de diphones répondant à ces critères sont disponibles par exemple auprès du Centre National d'Etudes des Télécommunications.

Les diphones sont alors utilisés dans un processus d'analyse et de synthèse schématisé sur la Figure 1. On décrira ce processus en supposant qu'il est mis en oeuvre dans un dispositif de synthèse ayant la constitution montrée en figure 2, destiné à être relié à un calculateur hôte, tel que le processeur central d'un ordinateur personnel. On supposera également que la fréquence d'échantillonnage donnant la représentation des diphones est de 16 kHz.

Le dispositif de synthèse (Figure 2) comporte alors une mémoire vive principale 16 qui contient un micro-programme de calcul, le dictionnaire de diphones 10 (c'est-à-dire des formes d'onde représentées par des échantillons) rangés dans l'ordre des adresses du descripteur, la table 12 constituant le descripteur de dictionnaire, et une fenêtre de Hanning, échantillonnée par exemple sur 500 points. La mémoire vive 16 constitue également mémoire de micro-trame et mémoire de travail. Elle est reliée par un bus de données 18 et un bus d'adresses 20 à un accès 22 au calculateur hôte.

Chaque micro-trame émise pour restituer un phonème (Figure 2) est constituée, pour chacun des deux phonèmes P et P + 1 qui interviennent
- du numéro d'ordre du phonème,
- de la valeur de la période au début du phonème, de la valeur de période à la fin du phonème, et
- de la durée totale du phonème pouvant être remplacée par la durée du diphone pour le second phonème.

Le dispositif comprend encore, reliés aux bus 18 et 20, une unité de calcul locale 24 et un circuit d'aiguillage 26. Ce dernier permet de relier une mémoire vive 28 servant de tampon de sortie soit vers le calculateur, soit vers un contrôleur 30 de convertisseur numérique/analogique 32 de sortie. Ce dernier attaque un filtre passe-bas 34, généralement limité à 8 kHz, qui alimente un amplificateur de parole 36.

Le fonctionnement du dispositif est le suivant.

Le calculateur hôte (non représenté) charge les micro-trames dans le tableau réservé en mémoire 16, par l'intermédiaire de l'accès 22 et des bus 18 et 20, puis il commande le début de synthèse à l'unité de calcul 24. Cette unité de calcul recherche le numéro du phonème courant P, du phonème suivant P + 1 et du phonème précédent P-1 dans le tableau de micro-trames, à l'aide d'un index mémorisé dans la mémoire de travail, initialisée à 1.

Dans le cas du premier phonème, l'unité de calcul vient chercher uniquement les numéros du phonème courant et du phonème suivant. Dans le cas du dernier phonème, elle vient chercher le numéro du phonème précédent et celui du phonème courant.

Dans le cas général, un phonème est constitué de deux demi-diphones ; l'adresse de chaque diphone est recherchée par adressage matriciel dans le descripteur du dictionnaire par la formule suivante :

numéro du descripteur de diphone =
numéro du 1er phonème + (numéro du 2ème phonème-1)*nombre de diphones

Sons voisés

L'unité de calcul charge, en mémoire de travail 16, l'adresse du diphone, sa longueur, son milieu ainsi que les trente-cinq marques de voisement. Elle charge ensuite, dans un tableau descripteur du phonème, les marques de voisement correspondant à la deuxième partie du diphone. Puis elle recherche, dans le dictionnaire de formes d'onde, la deuxième partie du diphone, qu'elle place dans un tableau représentant le signal du phonème d'analyse. Les marques conservées dans le tableau descripteur du phonème sont décrémentées de la valeur du milieu du diphone.

Cette opération est répétée pour la deuxième partie du phonème constituée par la première partie du deuxième diphone. Les marques de voisement de la première partie du deuxième diphone sont ajoutées aux marques de voisement du phonème et incrémentées de la valeur du milieu du phonème.

Dans le cas des sons voisés, l'unité de calcul, à partir des paramètres prosodiques (durée, période début et période fin du phonème) détermine alors le nombre de périodes nécessaire à la durée du phonème, suivant la formule :

nombre de périodes = 2*durée du phonème / (période début + période fin)

L'unité de calcul range en mémoire le nombre de marques du phonème naturel, égal au nombre de marques de voisement, puis détermine le nombre de périodes à éliminer ou à ajouter en effectuant la différence entre le nombre de périodes de synthèse et le nombre de périodes d'analyse, différence qui est fixée par la modification de tonalité à introduire à partir de celle qui correspond au dictionnaire.

Pour chaque période de synthèse retenue, l'unité de calcul détermine ensuite la période d'analyse retenue parmi les périodes du phonème à partir des considérations suivantes :

- la modification de la durée peut être considérée comme la mise en correspondance, par déformation de l'axe des temps du signal de synthèse, des n marques de voisement du signal d'analyse et des p marques du signal de synthèse, n et p étant des entiers prédéterminés ;
- à chacune des p marques du signal de synthèse doit être associée la marque la plus proche du signal d'analyse.

La duplication ou, au contraire, l'élimination de périodes également réparties sur tout le phonème modifie la durée de celui-ci.

Il faut noter qu'on n'aura pas à extraire une forme d'onde élémentaire à partir des deux périodes adjacentes de transition entre diphones : l'opération d'addition-recouvrement des fonctions élémentaires extraites des deux dernières périodes du premier diphone et des deux premières périodes du deuxième diphone permet le lissage entre ces diphones comme le montre la figure 5.

Pour chaque période de synthèse, l'unité de calcul détermine le nombre de points à ajouter ou à supprimer à la période d'analyse en effectuant la différence entre cette dernière et la période de synthèse.

Comme on l'a indiqué plus haut, il est avantageux de choisir la largeur de la fenêtre d'analyse de la façon suivante, illustrée en Figure 3 :

- si la période de synthèse est inférieure à la période d'analyse (lignes A et B de la Figure 3), la taille de la fenêtre 38 est le double de la période de synthèse ;
- dans le cas contraire, la taille de la fenêtre 40 est obtenue en multipliant par deux la plus faible des valeurs de la période d'analyse courante et de la période d'analyse précédente (lignes C et D).

L'unité de calcul détermine un pas d'avancement dans la lecture des valeurs de la fenêtre, tabulée par exemple sur 500 points, le pas étant alors égal à 500 divisé par la taille de la fenêtre précédemment calculée. Elle lit dans la mémoire tampon de signal du phonème d'analyse 28 les échantillons de la période précédente et de la période courante, les pondère par la valeur de la fenêtre de Hanning 38 ou 40 indexée par le numéro de l'échantillon courant multiplié par le pas d'avancement dans la fenêtre tabulée et ajoute, au fur et à mesure, les valeurs calculées à la mémoire tampon du signal de sortie indexé par la somme du compteur de l'échantillon courant de sortie et de l'index de recherche des échantillons du phonème d'analyse. Le compteur de sortie courant est ensuite incrémenté de la valeur de la période de synthèse.

Sons sourds (non voisés)

Pour les phonèmes sourds, le traitement est analogue au précédent, excepté que la valeur des

pseudo-périodes (distance entre deux marques de voisement) n'est jamais modifiée : l'élimination de pseudo-périodes au centre du phonème diminue simplement la durée de celui-ci.

On n'augmente pas la durée de phonèmes sourds, excepté par addition de zéros au milieu des phonèmes "silence".

Le fenêtrage s'effectue par période pour normaliser la somme des valeurs des fenêtres appliquées au signal :

- du début de la période précédente à la fin de la période précédente, le pas d'avancement dans la lecture de la fenêtre tabulée est (dans le cas d'une tabulation sur 500 points) égal à 500 divisé par deux fois la durée de la période précédente ;
- du début de la période courante à la fin de la période courante, le pas d'avancement dans la fenêtre tabulée est égal à 500 divisé par deux fois la durée de la période courante plus un décalage constant de 250 points.

A la fin du calcul du signal d'un phonème de synthèse, l'unité de calcul range la dernière période du phonème d'analyse et de synthèse dans la mémoire tampon 28 qui permet la transition entre phonèmes. Le compteur de l'échantillon courant de sortie est décrémenté de la valeur de la dernière période de synthèse.

Le signal ainsi généré est envoyé, par blocs de 2048 échantillons, dans un de deux espaces mémoire reservés à la communication entre l'unité de calcul et le contrôleur 30 du convertisseur numérique/analogique 32. Dès que le premier bloc est chargé dans la première zone tampon, le contrôleur 30 est activé par l'unité de calcul et vide cette première zone tampon. Pendant ce temps, l'unité de calcul remplit une deuxième zone tampon de 2048 échantillons. L'unité de calcul vient ensuite alternativement tester ces deux zones tampons grâce à un drapeau pour y charger le signal numérique de synthèse à la fin de chaque séquence de synthèse d'un phonème. Le contrôleur 30, en fin de lecture de chaque zone tampon, positionne le drapeau correspondant. En fin de synthèse, le contrôleur vide la dernière zone tampon et positionne un drapeau de fin de synthèse que le calculateur hôte peut lire via l'accès de communication 22.

L'exemple de spectre de signal de parole voisé d'analyse et de synthèse illustré en Figures 4A-4C montre que les transformations temporelles du signal numérique de parole n'affectent pas l'enveloppe du signal de synthèse, tout en modifiant la distance entre harmoniques, c'est-à-dire la fréquence fondamentale du signal de parole.

La complexité du calcul reste faible : le nombre d'opérations par échantillon est en moyenne de deux multiplications et deux additions pour la pondération et la sommation des fonctions élémentaires fournies par l'analyse.

L'invention est susceptible de nombreuses variantes de réalisation et, en particulier, comme on l'a indiqué plus haut, une fenêtre de largeur supérieure à deux périodes, comme le montre la Figure 6, éventuellement de taille fixe, peut donner des résultats acceptables.

On peut aussi utiliser le procédé de modification de la fréquence fondamentale sur des signaux numériques de parole en dehors de son application à la synthèse par diphones.

Revendications

1. Procédé de synthèse de parole à partir d'éléments sonores (mots, syllabes, diphones ...) caractérisé en ce que :
- on effectue, au moins sur les sons voisés des éléments sonores, une analyse par fenêtrage sensiblement centré sur le début de chaque réponse impulsionnelle du conduit vocal à l'excitation des cordes vocales à l'aide d'une fenêtre de filtrage présentant une amplitude décroissant jusqu'à zéro aux bords de la fenêtre dont la largeur est au moins égale à deux fois la période fondamentale d'origine ou deux fois la période fondamentale de synthèse,
- on replace les signaux résultant du fenêtrage correspondant à chaque élément sonore, avec un décalage temporel de ceux-ci égal à la période fondamentale de synthèse, inférieure ou supérieure à la période fondamentale d'origine, suivant l'information prosodique concernant la fréquence fondamentale de synthèse,
- on effectue la synthèse par sommation des signaux ainsi décalés.

2. Procédé de synthèse de parole selon la revendication 1, caractérisé en ce qu'on réalise un dictionnaire d'éléments sonores, par exemple de diphones, on découpe le texte à synthétiser en micro-trames identifiées chacune par le numéro de l'élément sonore correspondant (diphone) et au moins une information prosodique, constituée au moins par la valeur de la fréquence fondamentale en début et en fin d'élément et par la durée de l'élément.

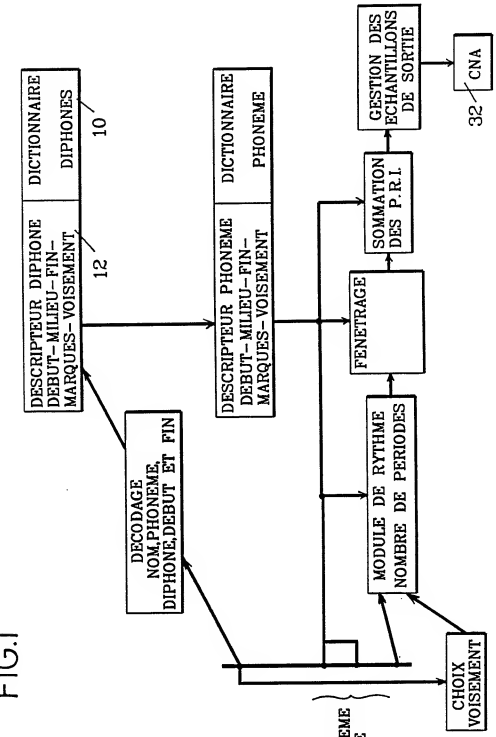
3. Procédé de synthèse de parole selon l'une des revendications 1 et 2, caractérisé en ce que la largeur de la fenêtre est égale à deux fois la période d'origine en cas de diminution de la fréquence fondamentale ou deux fois la période finale de synthèse en cas d'augmentation de la fréquence fondamentale.

4. Procédé de synthèse de parole selon l'une des revendications 1 à 3, caractérisé en ce que la fenêtre est une fenêtre de Hanning.

5. Dispositif de synthèse de parole par mise en

oeuvre du procédé selon la revendication 1, caractérisé en ce qu'il comprend, reliés à des bus (18,20) : une mémoire vive principale (16) qui contient un micro-programme de calcul, un dictionnaire de diphones (10) constitués de formes d'onde représentées par des échantillons rangés dans l'ordre des adresses d'un descripteur (12) de dictionnaire, et une fenêtre de Hanning échantillonnée, ladite mémoire vive (16) constituant également mémoire de micro-trame et mémoire de travail ; une unité de calcul locale (24) et un circuit d'aiguillage (26) permettant de relier une mémoire vive (28) servant de tampon de sortie soit vers l'unité de calcul, soit vers un contrôleur (30) de convertisseur numérique/analogique (32) de sortie attaquant un filtre passe-bas (34) qui alimente un amplificateur de parole (36).

FIG.1


DESCRIPTEUR DIPHONE
DEBUT-MILIEU-FIN-
MARQUES-VOISEMENT
12
DICTIONNAIRE
DIPHONES
10
DECODAGE
NOM PHONEME,
DIPHONE,DEBUT ET FIN
DESCRIPTEUR PHONEME
DEBUT-MILIEU-FIN-
MARQUES-VOISEMENT
DICTIONNAIRE
PHONEME
MODULE DE RYTHME
NOMBRE DE PERIODES
FENETRAGE
SOMMATION
DES P.R.I.
GESTION DES
ECHANTILLONS
DE SORTIE
32
CNA
CHOIX
VOISEMENT
NOM PHONEME P-1,P,P+1
PERIODE DEBUT DE PHONEME
PERIODE FIN DE PHONEME
DUREE PHONEME

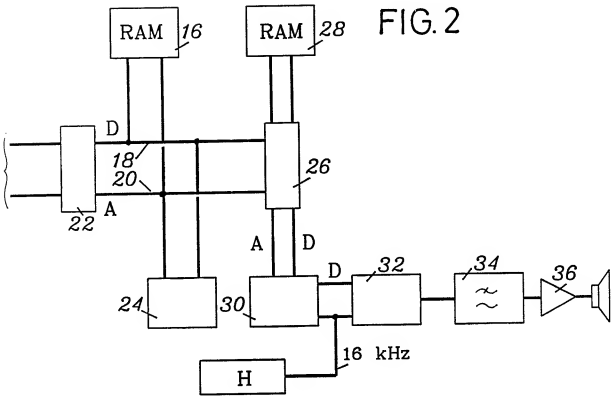

RAM
16
RAM
28
D
18
20
A
22
26
A
D
D
32
34
36
24
30
16 kHz
H

FIG. 2

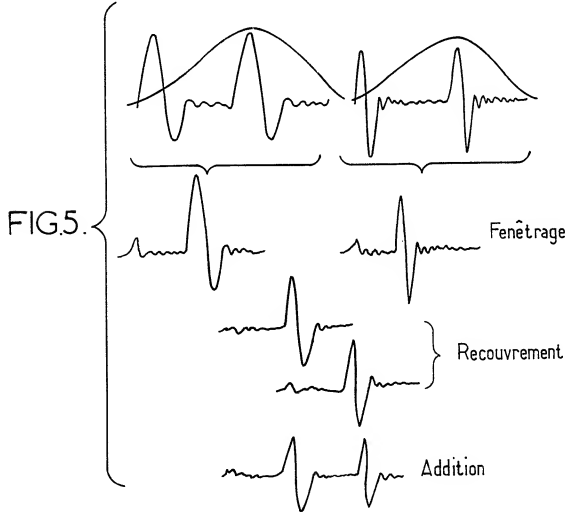

Fenêtrage
Recouvrement
Addition

FIG.5.

FIG. 3

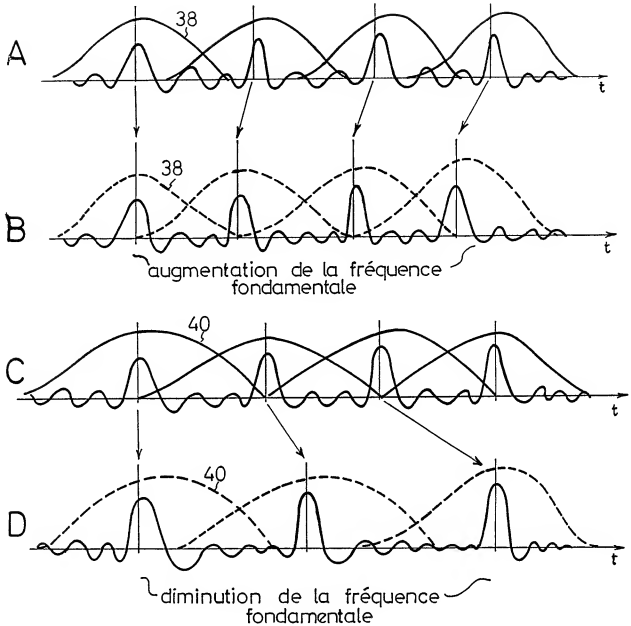

A
38
t
B
38
t
augmentation de la fréquence fondamentale
C
40
t
D
40
t
diminution de la fréquence fondamentale

FIG. 6

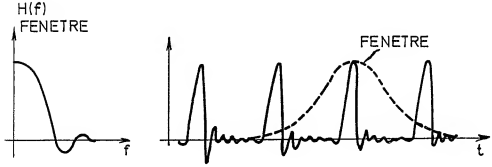

H(f)
FENETRE
f
FENETRE
t

FIG. 4A

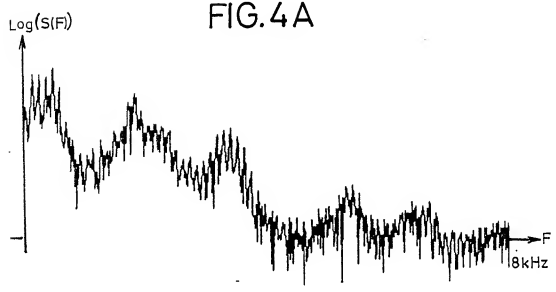


FIG. 4B

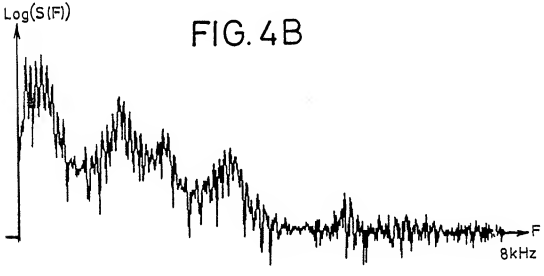
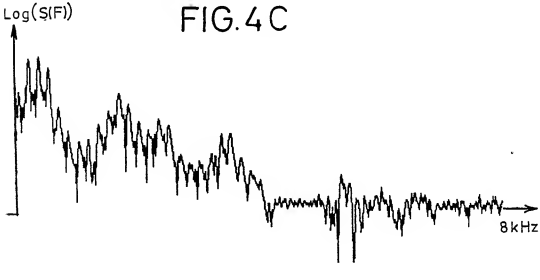


FIG. 4C



Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numero de la demande

EP 89 40 2394

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (Int. Cl.5)
D,X	ICASSP 86, (IEEE-IECEJ-ASJ INTERNATIONAL) CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Tokyo, 7-11 avril 1986, vol. 3, pages 2015-2018, IEEE, New York, US; F.J. CHARPENTIER et al.: "Diphone synthesis using an overlap-add technique for speech waveforms concatenation" * Figures 1,3 * ---	1,4	G 10 L 5/04
X	ICASSP 86, (IEEE-IECEJ-ASJ INTERNATIONAL) CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Tokyo, 7-11 avril 1986, vol. 3, pages 1705-1708, IEEE, New York, US; J. MAKHOUL et al.: "Time-scale modification in medium to low rate speech coding" * Paragraphe 2: "TSM of speech using SOLA"; page 1707, colonne de gauche: "Time-scale modification" * ---	1,3,4	
			DOMAINES TECHNIQUES RECHERCHES (Int. Cl.5)
A	IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING ,vol. ASSP-27, no. 2, avril 1979, pages 121-133, IEEE, New York, US; D. MALAH: "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals" * Page 121, colonne de droite, lignes 4-7 * --- -/-	1,3,4	G 10 L 5/04

Le présent rapport a été établi pour toutes les revendications

Lieu de la recherche	Date d'achèvement de la recherche	Examinateur
LA HAYE	20-12-1989	ARMSPACH J.F.A.M.

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P0402)

Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numero de la demande

EP 89 40 2394

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (Int. Cl.5)
D,A	ICASSP 87 - IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Dallas, 6-9 avril 1987, vol. 3, pages 1426-1429, IEEE, New York, US; K. LUKASZEWICZ et al.: "Microphonemic method of speech synthesis" * Figure 4 * ---	1	
A	ICASSP 82 - IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Paris, 3-5 mai 1982, vol. 3, pages 1597-1600, IEEE, New York, US; J.L. COURBON et al.: "Sparte: a text-to-speech machine using synthesis by diphones" * Page 1598, colonne de gauche, lignes 17-27 * -----	2	
			DOMAINES TECHNIQUES RECHERCHES (Int. Cl.5)

Le présent rapport a été établi pour toutes les revendications

Lieu de la recherche	Date d'achèvement de la recherche	Examinateur
LA HAYE	20-12-1989	ARMSPACH J.F.A.M.

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P0402)

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Numéro de publication : **0 363 233 B1**

(12) # FASCICULE DE BREVET EUROPEEN

(45) Date de publication du fascicule du brevet : **30.11.94 Bulletin 94/48**

(51) Int. Cl.[5] : **G10L 5/04**

(21) Numéro de dépôt : **89402394.4**

(22) Date de dépôt : **01.09.89**

(54) **Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'onde.**

(30) Priorité : **02.09.88 FR 8811517**

(43) Date de publication de la demande : **11.04.90 Bulletin 90/15**

(45) Mention de la délivrance du brevet : **30.11.94 Bulletin 94/48**

(84) Etats contractants désignés : **BE DE ES GB IT NL SE**

(56) Documents cités :
ICASSP 86, (IEEE-IECEJ-ASJ INTERNATIONAL) CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Tokyo, 7-11 avril 1986, vol. 3, pages 2015-2018, IEEE, New York, US; F.J. CHARPENTIER et al.: "Diphone synthesis using an overlap-add technique for speech waveforms concatenation"
ICASSP 86, (IEEE-IECEJ-ASJ INTERNATIONAL) CONFERENCE ON ACOUSTICS, SPEECH,AND SIGNAL PROCESSING, Tokyo, 7-11 avril 1986, vol. 3, pages 1705-1708, IEEE, New York, US; J. MAKHOUL et al.: "Time-scale modification in medium to low rate speech coding"

(56) Documents cités :
IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING ,vol. ASSP-27,no. 2, avril 1979, pages 121-133, IEEE, New York, US; D. MALAH: "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals"
ICASSP 87 - IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Dallas, 6-9 avril 1987, vol. 3, pages 1426-1429, IEEE, New York,US; K. LUKASZEWICZ et al.: "Microphonemic method of speech synthesis"
ICASSP 82 - IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, Paris, 3-5 mai 1982, vol. 3, pages 1597-1600, IEEE, New York, US;J.L. COURBON et al.: "Sparte: a text-to-speech machine using synthesis bydiphones"

(73) Titulaire : **FRANCE TELECOM**
6, Place d'Alleray
F-75015 Paris (FR)

(72) Inventeur : **Hamon, Christian**
Le Grand Clos Quevert
F-22100 Dinan (FR)

(74) Mandataire : **Fort, Jacques**
CABINET PLASSERAUD
84, rue d'Amsterdam
F-75009 Paris (FR)

EP 0 363 233 B1

Description

L'invention concerne les procédés et dispositifs de synthèse de la parole ; elle concerne, plus particulièrement, la synthèse à partir d'un dictionnaire d'éléments sonores par découpage du texte à synthétiser en microtrames identifiées chacune par un numéro d'ordre d'élément sonore correspondant et par des paramètres prosodiques (information de hauteur de son au début et à la fin de l'élément sonore et durée de l'élément sonore), puis par adaptation et concaténation des éléments sonores par une procédure d'addition-recouvrement.

Les éléments sonores stockés dans le dictionnaire seront fréquemment des diphones, c'est-à-dire des transitions entre phonèmes, ce qui permet, pour la langue française, de se contenter d'un dictionnaire d'environ 1300 éléments sonores ; on peut cependant utiliser des éléments sonores différents, par exemple des syllabes ou même des mots. Les paramètres prosodiques sont déterminés en fonction de critères portant sur le contexte : la hauteur de son qui correspond à l'intonation, dépend de l'emplacement de l'élément sonore dans un mot et dans la phrase et la durée donnée à l'élément sonore est fonction du rythme de la phrase.

Il faut rappeler au passage que les méthodes de synthèse de la parole se subdivisent en deux groupes. Celles qui utilisent un modèle mathématique du conduit vocal (synthèse par prédiction linéaire, synthèse à formants et synthèse à transformée de Fourier rapide) font intervenir une déconvolution de la source et de la fonction de transfert du conduit vocal et exigent en général une cinquantaine d'opérations arithmétiques par échantillon numérique de la parole avant conversion numérique-analogique et restitution.

Cette déconvolution source-conduit vocal permet d'une part la modification de la valeur de la fréquence fondamentale des sons voisés, c'est-à-dire des sons qui ont une structure harmonique et sont provoqués par vibration des cordes vocales, et d'autre part la compression des données représentant le signal de parole.

Celles qui appartiennent au second groupe de procédés utilisent la synthèse dans le domaine temporel par concaténation de formes d'onde. Cette solution a l'avantage de la flexibilité d'emploi et de la possibilité de réduire considérablement le nombre d'opérations arithmétiques par échantillons. En contrepartie, elle ne permet pas de réduire le débit nécessaire à la transmission autant que les méthodes basées sur un modèle mathématique. Mais cet inconvénient disparaît lorsqu'on recherche essentiellement une bonne qualité de restitution sans être gêné par la nécessité de transmettre des données sur un canal étroit.

La synthèse de parole suivant la présente invention appartient au second groupe. Elle trouve une application particulièrement importante dans le domaine de la transformation d'une chaîne orthographique (constituée par exemple par le texte fourni par une imprimante) en un signal de parole, par exemple restitué directement ou émis sur une ligne téléphonique normale.

On connait déjà (Diphone synthesis using an overlap-add technique for speech waveforms concatenation, CHARPENTIER et al, ICASSP 1986, IEEE-IECEJ-ASJ International Conference on Acoustics Speech and Signal Processing, pages 2 015-2 018)un procédé de synthèse de parole à partir d'éléments sonores utilisant une technique d'addition-recouvrement de signaux à court-terme. Mais il s'agit de signaux à court-terme de synthèse, avec normalisation du recouvrement des fenêtres de synthèse, obtenus par un processus très complexe :

- analyse du signal original par fenêtrage synchrone du voisement ;
- transformée de Fourier du signal à court-terme ;
- détection d'enveloppe ;
- homothétie de l'axe fréquentiel sur le spectre de la source ;
- pondération du spectre modifié de la source par l'enveloppe du signal d'origine ;
- transformée de Fourier inverse.

La présente invention vise notamment à fournir un procédé relativement simple et permettant une reproduction acceptable de la parole. Elle part de l'hypothèse qu'on peut considérer les sons voisés comme la somme des réponses impulsionnelles d'un filtre, stationnaire durant plusieurs millisecondes, (correspondant au conduit vocal) excité par une suite de Dirac, c'est-à-dire par un "peigne d'impulsions", de façon synchrone de la fréquence fondamentale de la source, c'est-à-dire des cordes vocales, ce qui se traduit dans le domaine spectral par un spectre harmonique, les harmoniques étant espacés de la fréquence fondamentale et pondérés par une enveloppe présentant des maxima appelés formants, dépendant de la fonction de transfert du conduit vocal.

On a déjà proposé (Micro-phonemic method of speech synthesis, Lucaszewic et al, ICASSP 1987, IEEE, pages 1426-1429) d'effectuer une synthèse de parole où la diminution de la fréquence fondamentale des sons voisés, lorsqu'elle est nécessaire pour respecter des données prosodiques, est effectuée par insertion de zéros, les microphonèmes stockés devant alors obligatoirement correspondre à la hauteur maximale possible du son à restituer, ou bien (brevet US 4 692 941) de diminuer de la même manière par insertion de zéros la fréquence fondamentale, et d'augmenter celle-ci en diminuant la taille de chaque période. Ces deux méthodes introduisent sur le signal de parole des distorsions non négligeables lors de la modification de la fréquence fondamentale.

La présente invention vise à fournir un procédé et un dispositif de synthèse à concaténation de formes d'onde ne présentant pas la limitation ci-dessus et permettant de fournir une parole de bonne qualité, tout en ne nécessitant qu'un faible volume de calculs arithmétiques.

Dans ce but, l'invention propose notamment un procédé suivant la revendication 1.

Ces opérations constituent la procédure de recouvrement puis addition des formes d'onde élémentaires obtenues par fenêtrage du signal de parole.

En général, on utilisera des éléments sonores constitués par des diphones.

La largeur de la fenêtre peut varier entre des valeurs inférieures et supérieures à deux fois la période d'origine. Dans l'exemple de mise en oeuvre qui sera décrit plus loin, la largeur de la fenêtre est choisie avantageusement égale à environ deux fois la période d'origine en cas d'augmentation de la période fondamentale ou environ deux fois la période finale de synthèse en cas d'augmentation de la fréquence fondamentale, afin de compenser partiellement les modifications d'énergie dues au changement de la fréquence fondamentale, non compensées par une normalisation possible de l'énergie, tenant compte de la contribution de chaque fenêtre à l'amplitude des échantillons du signal numérique de synthèse : dans le cas d'une diminution de la période fondamentale, la largeur de la fenêtre sera donc inférieure à deux fois la période fondamentale d'origine. Il est peu souhaitable de descendre au dessous de cette valeur.

Du fait qu'il est possible de modifier la valeur de la fréquence fondamentale dans les deux sens, les diphones sont mémorisés avec la fréquence fondamentale naturelle du locuteur.

Avec une fenêtre de durée égale à deux périodes fondamentales consécutives dans le cas voisé, on obtient des formes d'onde élémentaires dont le spectre représente sensiblement l'enveloppe du spectre du signal de parole ou spectre à court terme large bande -du fait que ce spectre est obtenu par convolution du spectre harmonique du signal de parole et de la réponse fréquentielle de la fenêtre, qui dans ce cas possède une largeur de bande supérieure à la distance entre harmoniques- ; la redistribution temporelle de ces formes d'onde élémentaires donnera un signal possédant sensiblement la même enveloppe que le signal d'origine mais une distance entre harmoniques modifiée.

Avec une fenêtre de durée supérieure à deux périodes fondamentales, on obtient des formes d'onde élémentaires dont le spectre est encore harmonique, ou spectre à court terme bande étroite -du fait que cette fois-ci la réponse fréquentielle de la fenêtre est moins large que la distance entre harmoniques- ; la redistribution temporelle de ces formes d'onde élémentaires donnera un signal possédant, comme le signal de synthèse précédent, sensiblement la même enveloppe que le signal d'origine à ceci près qu'on aura introduit des termes de réverbération (signaux dont le spectre possède une amplitude moindre, une phase différente, mais la même forme que le spectre d'amplitude du signal d'origine), dont l'effet ne sera audible qu'au delà de largeurs de fenêtre d'environ trois périodes, cet effet de réverbération ne dégradant pas la qualité du signal de synthèse lorsque son amplitude est faible.

On peut notamment utiliser une fenêtre de Hanning, bien que d'autres formes de fenêtre soient également acceptables.

Le traitement défini ci-dessus peut également être appliqué aux sons dits sourds ou non voisés, pouvant être représentés par un signal dont la forme s'apparente à celle d'un bruit blanc, mais sans synchronisation des signaux fenêtrés : ceci a pour but d'homogénéiser le traitement sur les sons sourds et les sons voisés, ce qui permet d'une part le lissage entre éléments sonores (diphones) et entre phonèmes sourds et voisés, et d'autre part une modification du rythme. Il se pose un problème à la jonction entre diphones. Une solution pour écarter cette difficulté consiste à omettre l'extraction de formes d'onde élémentaires à partir des deux périodes fondamentales adjacentes de transition entre diphones (dans le cas des sons sourds, les marques de voisement sont remplacées par des marques posées arbitrairement) : on pourra soit définir une troisième fonction d'onde élémentaire en calculant la moyenne des deux fonctions d'onde élémentaires extraites de part et d'autre du diphone, soit utiliser la procédure d'addition-recouvrement directement sur ces deux fonctions d'onde élémentaires.

L'invention sera mieux comprise à la lecture de la description qui suit d'un mode particulier de mise en oeuvre de l'invention, donné à titre d'exemple non limitatif. La description se réfère aux dessins qui l'accompagnent, dans lesquels :

- la Figure 1 est un graphe destiné à illustrer la synthèse de la parole par concaténation de diphones et modification des paramètres prosodiques dans le domaine temporel, conformément à l'invention ;
- la Figure 2 est un schéma synoptique montrant une constitution possible du dispositif de synthèse, implanté sur un calculateur hôte ;
- la Figure 3 montre, à titre d'exemple, comment on modifie les paramètres prosodiques d'un signal naturel, dans le cas d'un phonème particulier ;
- les Figures 4A, 4B et 4C sont des graphiques destinés à montrer des modifications spectrales apportées à des signaux de synthèse voisés, la Figure 4A montrant le spectre d'origine, la Figure 4B le spectre avec diminution de la fréquence fondamentale et la Figure 4C le spectre avec augmentation de cette fréquen-

ce ;
- la Figure 5 est un graphique montrant un principe d'atténuation des discontinuités entre diphones ;
- la Figure 6 est un schéma montrant le fenêtrage sur plus de deux périodes.

La synthèse d'un phonème est effectuée à partir de deux diphones stockés dans un dictionnaire, chaque phonème étant composé de deux demi-diphones. Le son "é" dans "période" par exemple sera obtenu à partir du second demi-diphone de "pai" et du premier demi-diphone de "air".

Un module de traduction orthographique phonétique et de calcul de la prosodie (qui ne fait pas partie de l'invention) fournit à un instant donné, des indications identifiant :
- le phonème à restituer, d'ordre P
- le phonème précédent, d'ordre P-1
- le phonème suivant, d'ordre P+1

et donnant la durée à affecter au phonème P ainsi que les périodes au début et à la fin (Figure 1).

Une première opération d'analyse, qui n'est pas modifiée par l'invention, consiste à déterminer, par décodage du nom des phonèmes et des indications prosodiques, les deux diphones retenus pour le phonème à utiliser et le voisement.

Tous les diphones disponibles (au nombre de 1300 par exemple) sont mémorisés dans un dictionnaire 10 muni d'une table constituant le descripteur 12 et contenant l'adresse du début de chaque diphone (en nombre de blocs de 256 octets) la longueur du diphone et le milieu du diphone (ces deux derniers paramètres étant exprimés en nombre d'échantillons à partir du début) et des marques de voisement repérant le début de la réponse du conduit vocal à l'excitation des cordes vocales dans le cas d'un son voisé (au nombre de 35 par exemple). Des dictionnaires de diphones répondant à ces critères sont disponibles par exemple auprès du Centre National d'Etudes des Télécommunications.

Les diphones sont alors utilisés dans un processus d'analyse et de synthèse schématisé sur la Figure 1. On décrira ce processus en supposant qu'il est mis en oeuvre dans un dispositif de synthèse ayant la constitution montrée en figure 2, destiné à être relié à un calculateur hôte, tel que le processeur central d'un ordinateur personnel. On supposera également que la fréquence d'échantillonnage donnant la représentation des diphones est de 16 kHz.

Le dispositif de synthèse (Figure 2) comporte alors une mémoire vive principale 16 qui contient un micro-programme de calcul, le dictionnaire de diphones 10 (c'est-à-dire des formes d'onde représentées par des échantillons) rangés dans l'ordre des adresses du descripteur, la table 12 constituant le descripteur de dictionnaire, et une fenêtre de Hanning, échantillonnée par exemple sur 500 points. La mémoire vive 16 constitue également mémoire de micro-trame et mémoire de travail. Elle est reliée par un bus de données 18 et un bus d'adresses 20 à un accès 22 au calculateur hôte.

Chaque micro-trame émise pour restituer un phonème (Figure 2) est constituée, pour chacun des deux phonèmes P et P+1 qui interviennent
- du numéro d'ordre du phonème,
- de la valeur de la période au début du phonème, de la valeur de période à la fin du phonème, et
- de la durée totale du phonème pouvant être remplacée par la durée du diphone pour le second phonème.

Le dispositif comprend encore, reliés aux bus 18 et 20, une unité de calcul locale 24 et un circuit d'aiguillage 26. Ce dernier permet de relier une mémoire vive 28 servant de tampon de sortie soit vers le calculateur, soit vers un contrôleur 30 de convertisseur numérique/analogique 32 de sortie. Ce dernier attaque un filtre passe-bas 34, généralement limité à 8 kHz, qui alimente un amplificateur de parole 36.

Le fonctionnement du dispositif est le suivant.

Le calculateur hôte (non représenté) charge les micro-trames dans le tableau réservé en mémoire 16, par l'intermédiaire de l'accès 22 et des bus 18 et 20, puis il commande le début de synthèse à l'unité de calcul 24. Cette unité de calcul recherche le numéro du phonème courant P, du phonème suivant P+1 et du phonème précédent P-1 dans le tableau de micro-trames, à l'aide d'un index mémorisé dans la mémoire de travail, initialisée à 1. Dans le cas du premier phonème, l'unité de calcul vient chercher uniquement les numéros du phonème courant et du phonème suivant. Dans le cas du dernier phonème, elle vient chercher le numéro du phonème précédent et celui du phonème courant.

Dans le cas général, un phonème est constitué de deux demi-diphones ; l'adresse de chaque diphone est recherchée par adressage matriciel dans le descripteur du dictionnaire par la formule suivante :

numéro du descripteur de diphone = numéro du 1er phonème + (numéro du 2ème phonème - 1) * nombre de diphones

Sons voisés

L'unité de calcul charge, en mémoire de travail 16, l'adresse du diphone, sa longueur, son milieu ainsi que les trente-cinq marques de voisement. Elle charge ensuite, dans un tableau descripteur du phonème, les marques de voisement correspondant à la deuxième partie du diphone. Puis elle recherche, dans le dictionnaire de formes d'onde, la deuxième partie du diphone, qu'elle place dans un tableau représentant le signal du phonème d'analyse. Les marques conservées dans le tableau descripteur du phonème sont décrémentées de la valeur du milieu du di-

phone.

Cette opération est répétée pour la deuxième partie du phonème constituée par la première partie du deuxième diphone. Les marques de voisement de la première partie du deuxième diphone sont ajoutées aux marques de voisement du phonème et incrémentées de la valeur du milieu du phonème.

Dans le cas des sons voisés, l'unité de calcul, à partir des paramètres prosodiques (durée, période début et période fin du phonème) détermine alors le nombre de périodes nécessaire à la durée du phonème, suivant la formule :

nombre de périodes = 2 * durée du phonème / (période début + période fin)

L'unité de calcul range en mémoire le nombre de marques du phonème naturel, égal au nombre de marques de voisement, puis détermine le nombre de périodes à éliminer ou à ajouter en effectuant la différence entre le nombre de périodes de synthèse et le nombre de périodes d'analyse, différence qui est fixée par la modification de tonalité à introduire à partir de celle qui correspond au dictionnaire.

Pour chaque période de synthèse retenue, l'unité de calcul détermine ensuite la période d'analyse retenue parmi les périodes du phonème à partir des considérations suivantes :

- la modification de la durée peut être considérée comme la mise en correspondance, par déformation de l'axe des temps du signal de synthèse, des n marques de voisement du signal d'analyse et des p marques du signal de synthèse, n et p étant des entiers prédéterminés ;
- à chacune des p marques du signal de synthèse doit être associée la marque la plus proche du signal d'analyse.

La duplication ou, au contraire, l'élimination de périodes également réparties sur tout le phonème modifie la durée de celui-ci.

Il faut noter qu'on n'aura pas à extraire une forme d'onde élémentaire à partir des deux périodes adjacentes de transition entre diphones : l'opération d'addition-recouvrement des fonctions élémentaires extraites des deux dernières périodes du premier diphone et des deux premières périodes du deuxième diphone permet le lissage entre ces diphones comme le montre la figure 5.

Pour chaque période de synthèse, l'unité de calcul détermine le nombre de points à ajouter ou à supprimer à la période d'analyse en effectuant la différence entre cette dernière et la période de synthèse.

Comme on l'a indiqué plus haut, il est avantageux de choisir la largeur de la fenêtre d'analyse de la façon suivante, illustrée en Figure 3 :

- si la période de synthèse est inférieure à la période d'analyse (lignes A et B de la Figure 3), la taille de la fenêtre 38 est le double de la période de synthèse ;
- dans le cas contraire, la taille de la fenêtre 40 est obtenue en multipliant par deux la plus faible des valeurs de la période d'analyse courante et de la période d'analyse précédente (lignes C et D).

L'unité de calcul détermine un pas d'avancement dans la lecture des valeurs de la fenêtre, tabulée par exemple sur 500 points, le pas étant alors égal à 500 divisé par la taille de la fenêtre précédemment calculée. Elle lit dans la mémoire tampon de signal du phonème d'analyse 28 les échantillons de la période précédente et de la période courante, les pondère par la valeur de la fenêtre de Hanning 38 ou 40 indexée par le numéro de l'échantillon courant multiplié par le pas d'avancement dans la fenêtre tabulée et ajoute, au fur et à mesure, les valeurs calculées à la mémoire tampon du signal de sortie indexé par la somme du compteur de l'échantillon courant de sortie et de l'index de recherche des échantillons du phonème d'analyse. Le compteur de sortie courant est ensuite incrémenté de la valeur de la période de synthèse.

Sons sourds (non voisés)

Pour les phonèmes sourds, le traitement est analogue au précédent, excepté que la valeur des pseudo-périodes (distance entre deux marques de voisement) n'est jamais modifiée : l'élimination de pseudo-périodes au centre du phonème diminue simplement la durée de celui-ci.

On n'augmente pas la durée de phonèmes sourds, excepté par addition de zéros au milieu des phonèmes "silence".

Le fenêtrage s'effectue par période pour normaliser la somme des valeurs des fenêtres appliquées au signal :

- du début de la période précédente à la fin de la période précédente, le pas d'avancement dans la lecture de la fenêtre tabulée est (dans le cas d'une tabulation sur 500 points) égal à 500 divisé par deux fois la durée de la période précédente ;
- du début de la période courante à la fin de la période courante, le pas d'avancement dans la fenêtre tabulée est égal à 500 divisé par deux fois la durée de la période courante plus un décalage constant de 250 points.

A la fin du calcul du signal d'un phonème de synthèse, l'unité de calcul range la dernière période du phonème d'analyse et de synthèse dans la mémoire tampon 28 qui permet la transition entre phonèmes. Le compteur de l'échantillon courant de sortie est décrémenté de la valeur de la dernière période de synthèse.

Le signal ainsi généré est envoyé, par blocs de 2048 échantillons, dans un de deux espaces mémoire reservés à la communication entre l'unité de calcul et le contrôleur 30 du convertisseur numérique/ana-

logique 32. Dès que le premier bloc est chargé dans la première zone tampon, le contrôleur 30 est activé par l'unité de calcul et vide cette première zone tampon. Pendant ce temps, l'unité de calcul remplit une deuxième zone tampon de 2048 échantillons. L'unité de calcul vient ensuite alternativement tester ces deux zones tampons grâce à un drapeau pour y charger le signal numérique de synthèse à la fin de chaque séquence de synthèse d'un phonème. Le contrôleur 30, en fin de lecture de chaque zone tampon, positionne le drapeau correspondant. En fin de synthèse, le contrôleur vide la dernière zone tampon et positionne un drapeau de fin de synthèse que le calculateur hôte peut lire via l'accès de communication 22.

L'exemple de spectre de signal de parole voisé d'analyse et de synthèse illustré en Figures 4A-4C montre que les transformations temporelles du signal numérique de parole n'affectent pas l'enveloppe du signal de synthèse, tout en modifiant la distance entre harmoniques, c'est-à-dire la fréquence fondamentale du signal de parole.

La complexité du calcul reste faible : le nombre d'opérations par échantillon est en moyenne de deux multiplications et deux additions pour la pondération et la sommation des fonctions élémentaires fournies par l'analyse.

L'invention est susceptible de nombreuses variantes de réalisation et, en particulier, comme on l'a indiqué plus haut, une fenêtre de largeur supérieure à deux périodes, comme le montre la Figure 6, éventuellement de taille fixe, peut donner des résultats acceptables.

On peut aussi utiliser le procédé de modification de la fréquence fondamentale sur des signaux numériques de parole en dehors de son application à la synthèse par diphones.

Revendications

1. Procédé de synthèse de parole à partir d'éléments sonores (mots, syllabes, diphones,...), suivant lequel :

   (a) on effectue, au moins sur les sons voisés des éléments sonores, une analyse en appliquant une fenêtre de filtrage synchrone de la fréquence fondamentale d'origine, sensiblement centrée sur le début de chaque réponse impulsionnelle du conduit vocal à l'excitation des cordes vocales, présentant une amplitude décroissant jusqu'à zéro aux bords de la fenêtre, dont la largeur est au moins égale à environ deux fois la période fondamentale d'origine ou environ deux fois la période fondamentale de synthèse, selon que la période fondamentale de synthèse est supérieure ou inférieure à la période fondamentale d'origine,

   (b) on replace les signaux résultant du fenêtrage correspondant à chaque élément sonore, avec un décalage temporel de ceux-ci égal à la période fondamentale de synthèse, suivant une information prosodique concernant la fréquence fondamentale de synthèse, et

   (c) on effectue la synthèse par sommation des signaux ainsi décalés,

   caractérisé en ce que le procédé ne comporte pas de transformation spectrale des signaux analysés, visant à modifier la fréquence fondamentale de ces signaux, entre les étapes (a) et (b).

2. Procédé de synthèse de parole selon la revendication 1, caractérisé en ce qu'on réalise un dictionnaire d'éléments sonores, par exemple de diphones, on découpe le texte à synthétiser en micro-trames identifiées chacune par le numéro de l'élément sonore correspondant (diphone) et au moins une information prosodique, constituée au moins par la valeur de la fréquence fondamentale en début et en fin d'élément et par la durée de l'élément.

3. Procédé de synthèse de parole selon l'une des revendications 1 et 2, caractérisé en ce que la largeur de la fenêtre est égale à deux fois la période d'origine en cas de diminution de la fréquence fondamentale ou deux fois la période finale de synthèse en cas d'augmentation de la fréquence fondamentale.

4. Procédé de synthèse de parole selon l'une des revendications 1 à 3, caractérisé en ce que la fenêtre est une fenêtre de Hanning.

5. Dispositif de synthèse de parole executant le procédé selon la revendication 1, caractérisé en ce qu'il comprend, reliés à des bus (18,20) : une mémoire vive principale (16) qui contient un microprogramme de calcul, un dictionnaire de diphones (10) constitués de formes d'onde représentées par des échantillons rangés dans l'ordre des adresses d'un descripteur (12) de dictionnaire, et une fenêtre de Hanning échantillonnée, ladite mémoire vive (16) constituant également mémoire de micro-trame et mémoire de travail ; une unité de calcul locale (24) et un circuit d'aiguillage (26) permettant de relier une mémoire vive (28) servant de tampon de sortie soit vers l'unité de calcul, soit vers un contrôleur (30) de convertisseur numérique/analogique (32) de sortie attaquant un filtre passe-bas (34) qui alimente un amplificateur de parole (36).

Patentansprüche

1. Verfahren zur Sprachsynthese aus akustischen Elementen (Worten, Silben, Diphonen, ...) gemäß welchem:
   (a) bei wenigstens den stimmhaften Lauten der akustischen Elemente eine Analyse unter Anwendung eines Fensters zum synchronen Filtern der Ursprungsgrundfrequenz durchgeführt wird, welches im wesentlichen auf dem Anfang jeder Impulsantwort des Stimmkanals bei Anregung der Stimmbänder zentriert ist, welche eine an den Rändern des Fensters bis auf Null absinkende Amplitude aufweist, dessen Breite wenigstens ungefähr das Zweifache der Ursprungsgrundperiode oder ungefähr das Zweifache der Synthesegrundperiode ist, je nachdem, ob die Synthesegrundperiode größer oder kleiner als die Ursprungsgrundperiode ist,
   (b) die aus der Anwendung des Fensters resultierenden, jedem akustischen Element entsprechenden Signale mit einer zeitlichen Verschiebung derselben, welche gleich der Grundperiode der Synthese ist, gemäß einer die Grundfrequenz der Synthese betreffenden prosodischen Information wiederaufgestellt werden und
   (c) die Synthese durch Summierung der derart verschobenen Signale durchgeführt wird,
   dadurch gekennzeichnet,
   daß das Verfahren keine spektrale Transformation der analysierten Signale zwischen den Schritten (a) und (b) umfaßt, welche darauf abzielt, die Grundfrequenz dieser Signale zu modifizieren.

2. Verfahren zur Sprachsynthese nach Anspruch 1, dadurch gekennzeichnet, daß ein Lexikon von akustischen Elementen, z.B. von Diphonen, erstellt wird, der zu synthetisierende Text in Mikroensembles aufgeteilt wird, welche jeweils durch die Nummer des entsprechenden akustischen Elements (Diphon) und wenigstens eine prosodische, wenigstens von dem Wert der Grundfrequenz am Anfang und am Ende des Elements und von der Dauer des Elements gebildete Information identifiziert werden.

3. Verfahren zur Sprachsynthese nach einem der Ansprüche 1 und 2, dadurch gekennzeichnet, daß die Breite des Fensters das Zweifache der Ursprungsperiode im Fall der Verminderung der Grundfrequenz oder das Zweifache der Endperiode der Synthese im Fall der Verstärkung der Grundfrequenz ist.

4. Verfahren zur Sprachsynthese nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß das Fenster ein Hanning-Fenster ist.

5. Sprachsynthesevorrichtung zur Durchführung des Verfahrens nach Anspruch 1, dadurch gekennzeichnet, daß sie an Bussen (18, 20) angeschlossen umfaßt: ein Haupt-RAM (16), welches ein Berechnungs-Mikroprogramm, ein Lexikonterbuch (10) von Diphonen, welche von Wellenformen gebildet sind, die von in der Reihenfolge der Adressen eines Deskriptors (12) des Lexikons abgespeicherten Abtastwerten dargestellt werden, und ein abgetastetes Hanning-Fenster umfaßt, wobei das RAM (16) auch den Mikroensemblespeicher und den Arbeitsspeicher bildet; eine lokale Recheneinheit (24) und eine Verzweigungsschaltung (26), welche es ermöglicht, ein als Ausgangspuffer dienendes RAM (28) entweder mit der Recheneinheit zu verbinden oder mit einer Steuereinheit (30) eines Digital/Analog-Ausgangswandlers (32), welcher in ein einen Sprachverstärker (36) speisendes Tiefpaßfilter (34) mündet.

Claims

1. Method of speech synthesis from sound elements (words, syllabes, diphones,...), wherein:
   (a) analysis is carried out, at least on the voiced sounds of the sound elements, by windowing by means of a filtering window approximately centered on the beginning of each pulse response of the vocal tract to an excitation of the vocal cords, the window having an amplitude decreasing to zero at the edges of the window, whose width is at least equal to twice the original fundemental period or twice the fundamental synthesis period,
   (b) the signals resulting from windowing corresponding to each sound element are replaced with a time shift thereof equal to a fundamental synthesis period, which is lesser than or greater than the original fundamental period, responsive to prosodic information relating to the fundamental sysnthesis frequency,
   (c) synthesis is carried out by summing the thus shifted signals,
   characterized in that the method does not include a spectral transformation of the analysed signals, for modifying the fundamental frequency of said analysed signals, between steps (a) and (b).

2. Method of speech synthesis according to claim 1, characterized in that a dictionary of sound elements, for example diphones, is formed; the text to be synthesized is split into microframes

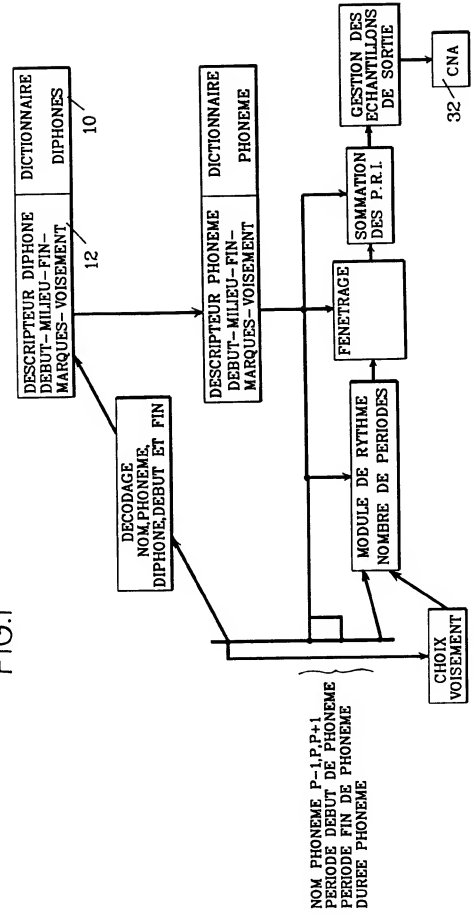
each identified by the serial number of the corresponding sound element (diphone) and at least one prosodic information, formed at least by the value of the fundamental frequency at the beginning and at the end of an element and by the duration of the element.

3. Method of speech synthesis according to any one of claims 1 and 2,
characterized in that the width of the window is equal to twice the original period in the case of reduction of the fundamental frequency or twice the final synthesis period in the case of increase of the fundamental frequency.

4. Method of speech synthesis according to any one of claims 1-3,
characterized in that the window is a Hanning window.

5. Device for speech synthesis carrying out the method of claim 1,
characterized in that it comprises, connected to buses (18, 20): a main random access memory (16) which contains a computing microprogram, a dictionary of diphones (10) formed of waveforms represented by samples stored in the order of the addresses of a dictionary descriptor (12) and a sampled Hanning window, said random access memory (16) also forming a microframe memory and a working memory; a local computing unit (24) and a routing circuit (26) making it possible to connect a random access memory (28) serving as output buffer either to the computing unit or to a controller (30) of an output digital/analog converter (32) driving a low pass filter (34) which feeds a speech amplifier (36).

FIG.1


DESCRIPTEUR DIPHONE
DEBUT-MILIEU-FIN-
MARQUES-VOISEMENT
DICTIONNAIRE
DIPHONES
10
12
DESCRIPTEUR PHONEME
DEBUT-MILIEU-FIN-
MARQUES-VOISEMENT
DICTIONNAIRE
PHONEME
DECODAGE
NOM,PHONEME,
DIPHONE,DEBUT ET FIN
MODULE DE RYTHME
NOMBRE DE PERIODES
FENETRAGE
SOMMATION
DES P.R.I.
GESTION DES
ECHANTILLONS
DE SORTIE
32
CNA
CHOIX
VOISEMENT
NOM PHONEME P-1,P,P+1
PERIODE DEBUT DE PHONEME
PERIODE FIN DE PHONEME
DUREE PHONEME

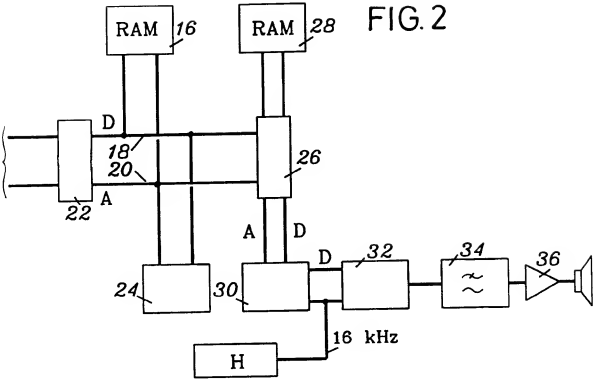

RAM
16
RAM
28
D
18
20
A
22
26
A
D
D
32
34
36
24
30
16 kHz
H

FIG. 2

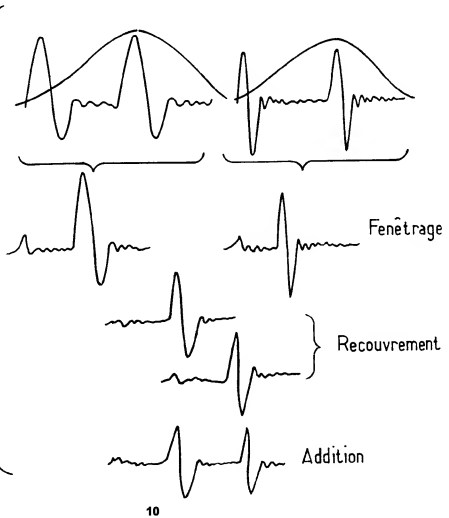

Fenêtrage
Recouvrement
Addition

FIG.5.

FIG. 3

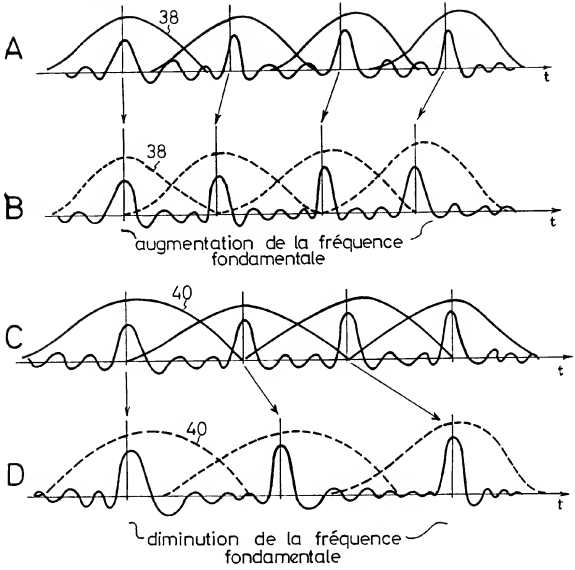

A
38
t
B
38
t
augmentation de la fréquence fondamentale
C
40
t
D
40
t
diminution de la fréquence fondamentale

FIG. 6

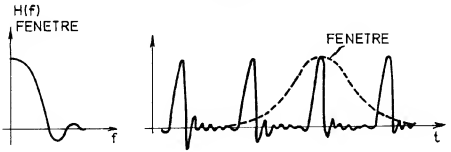

H(f)
FENETRE
f
FENETRE
t

FIG. 4A

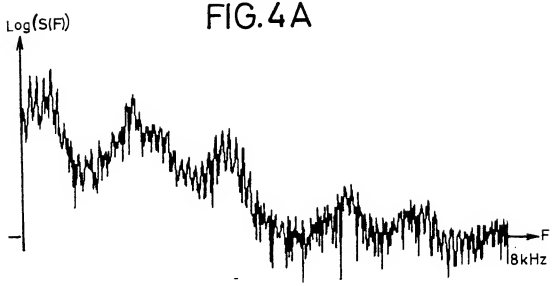


FIG. 4B

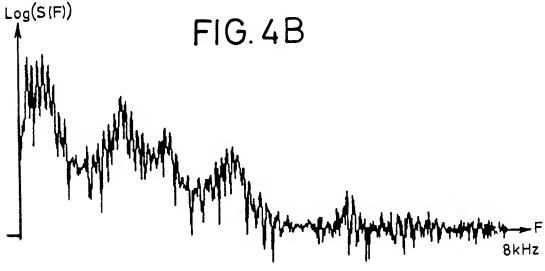


FIG. 4C